Heterogeneous Domain Adaptation and Classification by Exploiting the Correlation Subspace

Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang

***Abstract*— We present a novel domain adaptation approach for solving cross-domain pattern recognition problems, i.e., the data or features to be processed and recognized are collected from different domains of interest. Inspired by canonical correlation analysis (CCA), we utilize the derived correlation subspace as a joint representation for associating data across different domains, and we advance reduced kernel techniques for kernel CCA (KCCA) if nonlinear correlation subspace are desirable. Such techniques not only makes KCCA computationally more efficient, potential over-fitting problems can be alleviated as well. Instead of directly performing recognition in the derived CCA subspace (as prior CCA-based domain adaptation methods did), we advocate the exploitation of domain transfer ability in this subspace, in which each dimension has a unique capability in associating cross-domain data. In particular, we propose a novel support vector machine (SVM) with a correlation regularizer, named correlation-transfer SVM, which incorporates the domain adaptation ability into classifier design for cross-domain recognition. We show that our proposed domain adaptation and classification approach can be successfully applied to a variety of cross-domain recognition tasks such as cross-view action recognition, handwritten digit recognition with different features, and image-to-text or text-to-image classification. From our empirical results, we verify that our proposed method outperforms state-of-the-art domain adaptation approaches in terms of recognition performance.**

***Index Terms*— Canonical correlation analysis, domain adaptation, reduced kernels, support vector machine.**

I. Introduction

RESEARCHERS in image processing and computer vision communities have shown significant progresses on active issues like detection and recognition of objects, actions, or events in images and videos. The commonly underlying assumption is that both training and test data for addressing the above tasks exhibit the same distribution or are drawn from the same feature spaces (i.e., the same feature domains). When the distribution of the data changes, or when the feature domains for training and test data are different, one cannot expect the resulting recognition performance to be satisfactory. This is because the models learned with data in one domain would fail to predict the test data in another. Unfortunately, in many real-world applications, re-collecting training examples in different domains and observing the corresponding learning models might not be practical. Take action recognition using surveillance cameras for example, one typically extracts the features and designs the classifier using training data captured by one camera. However, the observed features and classification models might not generalize well if the test data are captured by another camera (e.g., in a different view).

Since re-collecting training data in different domains of interest is not practical, *transfer learning* aims at transferring the knowledge of the model learned from the source domain to that in the target domain, so that the unseen test data in the target domain can be predicted accordingly [1]. Among various scenarios of transfer learning, *domain adaptation* addresses the problem in which the *same* learning tasks need to be solved across different domains [2]–[9]. Typically, one has labeled data in the source domain for training, few or no labeled data are available in the target domain, but unseen test inputs in the target domain need to be recognized or retrieved. For example, recognizing actions across different camera views, determining the identity of input sketch images using photo ones as the gallery set, or retrieving relevant images using text as queries can all be considered as domain adaptation problems. To solve such problems, one generally needs to design learning models in the target domain (or in a common feature space) by leveraging labeled data observed in the source domain. Nevertheless, the challenge of domain adaptation comes from the fact that the source and target domain data are either drawn from different distributions of the same feature space [2], [10], [11], or the features for data in different domains are completely different [3]–[6], [8].

A. Our Contributions

In this paper, we present a novel domain adaptation framework for cross-domain recognition. We advance canonical correlation analysis (CCA) [12] for deriving a joint feature space for associating cross-domain data, and we propose a novel support vector machine (SVM) algorithm which incorporates the domain adaptation ability observed in the derived subspace

Manuscript received April 2, 2013; revised September 1, 2013 and December 30, 2013; accepted February 18, 2014. Date of publication March 11, 2014; date of current version March 31, 2014. This work was supported by the National Science Council of Taiwan under Grants NSC102-2221-E-001-005-MY2 and NSC103-2218-E-034-001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Andrea Cavallaro.

Y.-R. Yeh is with the Department of Applied Mathematics, Chinese Culture University, Taipei 11114, Taiwan (e-mail: yyr2@ulive.pccu.edu.tw).

C.-H. Huang is with the Technische Universität Müchen, Munich 85748, Germany (e-mail: huangc@in.tum.de).

Y.-C. F. Wang is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan (e-mail: ycwang@citi.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2310992

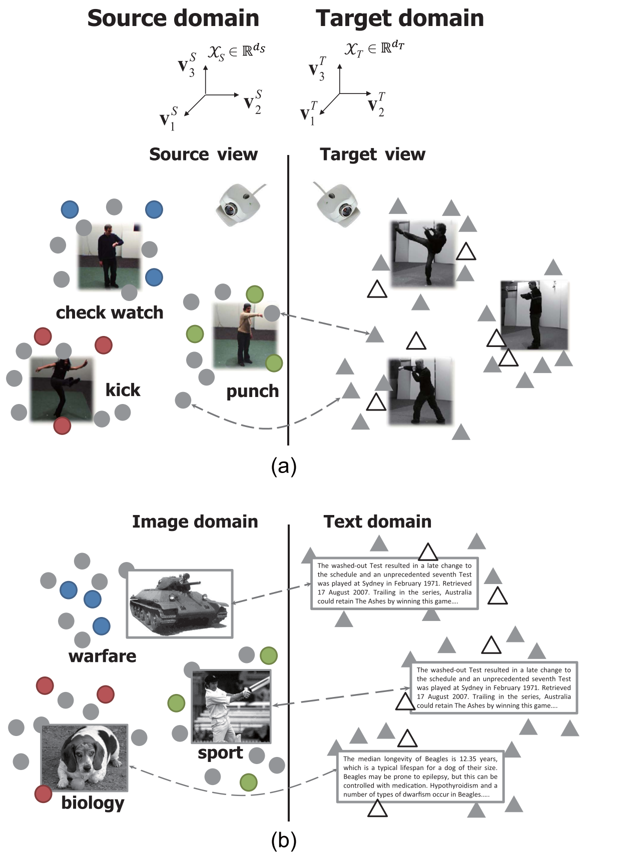


Fig. 1. Our approach of domain adaptation and two application. Note that instances in circles and triangles are data in the source and the target domain, respectively. We aim at utilizing labeled training data (colored circles) in the source domain and unlabeled data pairs (in gray) from both domains for recognizing testing data (in white) in the target domain. (a) Cross-camera action recognition. (b) Cross-modal classification.

for improved recognition. As shown in Fig. 1, we consider the scenarios in which only labeled data in the source domain (i.e., those shown in different colors in Fig. 1) and *unlabeled* data pairs (i.e., those shown in gray in Fig. 1) across different domains are available during training.

The above unlabeled data pairs are utilized to observe the joint feature space for domain adaptation. As discussed in Section III, we construct a shared feature representation in correlation subspaces as the domain adaptation model, and we exploit the domain transfer ability in such subspaces when designing the associated classifiers. In our work, we apply unlabeled cross-domain data pairs to derive the aforementioned feature representation for domain adaptation. In practice, unlabeled data pairs collected from different domains can be actions captured by different cameras (as considered in [6]–[8]), objects observed in different lighting, resolution, etc. conditions (e.g., [3]–[5]), or text and image portions presented in the same webpages (e.g., [13]). For example, to address the domain adaptation problem of image-to-text recognition, one might have labeled text data available for training and just the test data to be classified in the image domain. Since collecting and labeling the image data of interest (for training purposes) might be computationally expensive, it would be much easier

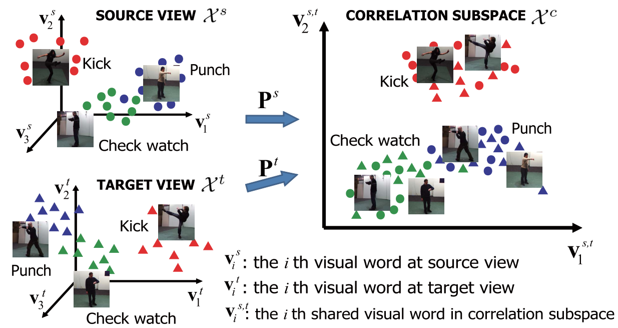


Fig. 2. Domain adaptation via CCA [2]. Note that $\mathbf{P}^s$ and $\mathbf{P}^t$ are the projection matrices produced by CCA, which maximize the correlation between projected data in the derived subspace $\mathcal{X}^c$.

for extracting text-image data pairs from webpages which need *not* belong to the same categories to be recognized. In other words, such unlabeled cross-domain data pairs can be easily applied to associate different data domains, so that the transfer of the observed classification model can be performed. Take Fig. 2 for example, once the common feature space is constructed by the aforementioned cross-domain data, one can simply project source-domain labeled and target-domain test data of interest into this space for recognition.

It is worth noting that, since the shared feature representation for domain adaptation using standard CCA might not be sufficient for practical cross-domain classification tasks, we further consider kernel CCA (KCCA) for deriving the nonlinear correlation subspace. In order to alleviate possible overfitting problems using KCCA, the *reduced kernel* technique [14] is applied to avoid possible trivial solution and potential over-fitting problems. In addition, the reduced kernel approach also allows one to reduce the storage and computational costs, since only a portion of the kernel matrix is required when implementing KCCA.

As for our proposed SVM model for cross-domain recognition, we advocate the exploitation of the *domain adaptation ability* in the resulting correlation subspace. As detailed later in Section IV, we propose a novel linear SVM (i.e., correlation-transfer SVM) which incorporates the domain adaptation ability when separating data from different categories using the associated correlation information. With the ability to bridge source and target domains when designing the classification model, our method is expected to achieve improved classification performance. Finally, we note that our proposed domain adaptation and classification algorithms can be regarded as a *heterogeneous domain adaptation* framework. This is because our framework is able to deal with more general transfer learning cases in which the distributions, feature domains, or feature dimensions in source and target domains are different (e.g., image vs. text in Fig. 1(b)). Later in our experiments, the effectiveness and robustness of our proposed method will be verified.

II. Related Work

Generally, methods to address domain adaptation problem can be divided into two categories: instance and feature-based approaches. We now briefly review them as follows.

A. Instance-Based Approaches

Instance-based approaches focus on source and target domain data with the same feature representation but drawn from different distributions. Its purpose is to estimate the importance of each source domain data instance, so that it can be re-weighted accordingly when constructing the learning model for predicting the target domain data. For example, Huang *et al.* [15] proposed a kernel-mean matching (KMM) algorithm, which observes the importance of source domain data by matching the means between the source and target domain data in a reproducing-kernel Hilbert space (RKHS). To alleviate possible biases when performing cross-validation across domains, Sugiyama *et al.* [16] proposed a unified framework which not only learns such importance via the minimization of Kullback-Leibler divergence but also cross-validates the estimated data in the target domain. Nevertheless, the methods of this category work on the premise that the source and target domain data are presented in the same feature space and thus the same feature dimension.

B. Feature-Based Approaches

Different from instance-based approaches which favor only cross-domain data in the same feature space, feature-based approaches aim at deriving a joint feature representation (e.g., a manifold or subspace) across domains. Many visual domain adaptation methods (including ours) are within this category. For example, methods based on the extraction of spatio-temporal descriptors and the construction of bag-of-words models [17], [18] have been proposed to recognize actions in videos. Gopalan *et al.* [4] characterized the source and target domain/view data as two points on a Grassmann manifold, and proposed to sample points between them (along the geodesic) as the shared feature representation for cross-camera action recognition. Similarly, Li and Zickler [8] derived discriminative intermediate virtual views by maximizing mutual information, and the shared representation can be applied to learn and predict unseen actions at the target view. Besides exploring the Grassmann manifold, Farhadi and Tabrizi [6] proposed to learn split-based features from source-view video frames based on local data structure. They converted such features to the corresponding frames at the target view, so that actions at the target view can be encoded and recognized accordingly. Thus, the assumption of their approach is that the local data structures in two domains are consistent. Saenko *et al.* [3] advocated a metric learning framework to model the visual domain shift (e.g., changes in resolution, lighting, viewpoint, etc). In [19], they further extended their work, so that source and target domain data in different dimensional feature spaces can be handled.

Recently, Liu *et al.* [7] advocated to construct a bilingual codebook as a shared feature representation for source and target domain data. With unlabeled data collected from both domains, their approach learned a shared codebook for action data captured by two cameras using a bipartite graph, and the bilingual words were obtained by spectral clustering. Although this approach does not require similarities of local data structure for data in different domains and allows feature dimensions to be different, the derived feature attributes in the joint representation are considered to be *equally important*, which may not be preferable if the (shared) features extracted from each domain have uncoordinated contributions. On the other hand, Duan et al. [20] proposed a joint optimization framework for integrating CCA and SVM. Although they achieved promising results due to the joint learning of the feature subspace and classifiers, they required *labeled* target-domain data during training, which might not be applicable for practical domain adaptation problems (like ours).

III. Domain Adaptation via Canonical Correlation Analysis

A. Learning Correlation Subspace via CCA or KCCA

The idea of domain adaptation for solving cross-domain classification tasks is to determine a common representation (e.g., a joint subspace) for features extracted from source and target views, so that the model trained in the source domain can be applied to recognize test data in the target domain. Among existing methods, canonical correlation analysis (CCA) is a very effective technique. It aims at maximizing the correlation between two variable sets [12], [21], and thus the derived subspace can be applied as the common feature representation for solving cross-domain classification problems.

For the sake of completeness, we briefly review CCA as follows. Given two sets of n centered unlabeled observations $\mathbf{X}^s = [\mathbf{x}_1^s, \ldots, \mathbf{x}_n^s] \in \mathbb{R}^{d_s \times n}$ and $\mathbf{X}^t = [\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t] \in \mathbb{R}^{d_t \times n}$ ($\mathbf{x}_i^s \in \mathcal{D}_u^s$ and $\mathbf{x}_i^t \in \mathcal{D}_u^t$) in source and target views respectively, CCA learns the projection vectors $\mathbf{u}^s \in \mathbb{R}^{d_s}$ and $\mathbf{u}^t \in \mathbb{R}^{d_t}$, which maximizes the correlation coefficient ρ:

$$\max_{\mathbf{u}^s, \mathbf{u}^t} \rho = \frac{\mathbf{u}^{s\top} \Sigma_{st} \mathbf{u}^t}{\sqrt{\mathbf{u}^{s\top} \Sigma_{ss} \mathbf{u}^s} \sqrt{\mathbf{u}^{t\top} \Sigma_{tt} \mathbf{u}^t}}, \tag{1}$$

where $\Sigma_{st} = \mathbf{X}^s \mathbf{X}^{t\top}$, $\Sigma_{ss} = \mathbf{X}^s \mathbf{X}^{s\top}$, $\Sigma_{tt} = \mathbf{X}^t \mathbf{X}^{t\top}$, and $\rho \in [0, 1]$. As proved in [21], $\mathbf{u}^s$ in (1) can be solved by a generalized eigenvalue decomposition problem:

$$\Sigma_{st} (\Sigma_{tt})^{-1} \Sigma_{st}^{\top} \mathbf{u}^s = \eta \Sigma_{ss} \mathbf{u}^s. \tag{2}$$

Once $\mathbf{u}^s$ is obtained, $\mathbf{u}^t$ can be calculated by $\Sigma_{tt}^{-1} \Sigma_{st} \mathbf{u}^s / \eta$, which has been dervied in [21]. In practice, regularization terms $\lambda_s \mathbf{I}$ and $\lambda_t \mathbf{I}$ need to be added into Σ_{ss} and Σ_{tt} to avoid overfitting and singularity problems. As a result, one solves the following problem instead:

$$\Sigma_{st} (\Sigma_{tt} + \lambda_t \mathbf{I})^{-1} \Sigma_{st}^{\top} \mathbf{u}^s = \eta (\Sigma_{ss} + \lambda_s \mathbf{I}) \mathbf{u}^s. \tag{3}$$

Generally, one can derive more than one pair of projection vectors $\{\mathbf{u}_i^s\}_{i=1}^d$ and $\{\mathbf{u}_i^t\}_{i=1}^d$ with corresponding ρ_i in a descending order (i.e., $\rho_i > \rho_{i+1}$). Note that d is the dimension number of the resulting CCA subspace, which is bounded by the minimum value of d_s and d_t (i.e., the dimension number of the CCA subspace is bounded by the minimum feature dimension of the data in either domain). As a result, the source and target view data ($\mathbf{X}^s$ and $\mathbf{X}^t$) can be projected onto this *correlation subspace* $\mathcal{X}^c \in \mathbb{R}^d$. Fig. 2 shows a CCA example for cross-view action recognition. Given data of three action classes in source and target views ($\mathcal{X}^s$ and $\mathcal{X}^t$), CCA determines projection matrices $\mathbf{P}^s = [\mathbf{u}_1^s, \ldots, \mathbf{u}_d^s] \in \mathbb{R}^{d_s \times d}$ and $\mathbf{P}^t = [\mathbf{u}_1^t, \ldots, \mathbf{u}_d^t] \in \mathbb{R}^{d_t \times d}$. Once the correlation subspace $\mathcal{X}^c \in \mathbb{R}^d$ spanned by $\{\mathbf{u}_i^{s,t}\}_{i=1}^d$ is derived, unseen test data at the target view can be directly recognized by the model trained from the source view data projected onto $\mathcal{X}^c$.

Similar to SVM, the use of linear models for CCA might not be sufficient in describing data when deriving the feature subspace. To overcome this problem, nonlinear mapping via kernel tricks can be applied to CCA and result in kernel CCA (KCCA) [21]. Let ϕ be a nonlinear mapping from $\mathcal{X}$ to a feature space $\mathcal{F}$, and the kernel function $k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$. It also has been shown that the standard CCA projection vectors $\mathbf{u}^s$ and $\mathbf{u}^t$ can represented by the linear combination of the data $\mathbf{X}^s$ and $\mathbf{X}^t$ [21]:

$$\mathbf{u}^s = \mathbf{X}^s \alpha^s \text{ and } \mathbf{u}^t = \mathbf{X}^t \alpha^t, \tag{4}$$

where $\alpha^s \in \mathbb{R}^n$ and $\alpha^t \in \mathbb{R}^n$ are weighting vectors for $\mathcal{F}^s$ and $\mathcal{F}^t$, respectively. Using the kernel trick, KCCA can be formulated as follows:

$$\max_{\alpha^s, \alpha^t} \rho = \frac{\alpha^{s\top} \mathbf{K}_s \mathbf{K}_t \alpha^t}{\sqrt{\alpha^{s\top} \mathbf{K}_s^2 \alpha^s} \sqrt{\alpha^{t\top} \mathbf{K}_t^2 \alpha^t}}, \tag{5}$$

where $\mathbf{K}_s = k(\mathbf{X}^s, \mathbf{X}^s)$ and $\mathbf{K}_t = k(\mathbf{X}^t, \mathbf{X}^t)$ are the kernel matrices corresponding to the source and target data. As shown in [21], deriving the solution to (5) is equivalent to solving the following problem:

$$\mathbf{K}_s \mathbf{K}_t \mathbf{K}_t^{-1} \mathbf{K}_s \alpha^s = \lambda^2 \mathbf{K}_s \mathbf{K}_s \alpha^s$$

or

$$\mathbf{K}_s \mathbf{K}_s \alpha^s = \lambda^2 \mathbf{K}_s \mathbf{K}_s \alpha^s$$

or

$$\mathbf{I} \alpha^s = \lambda^2 \alpha^s. \tag{6}$$

Similar to standard CCA, α^t can be determined by $\frac{\mathbf{K}_t^{-1} \mathbf{K}_s \alpha^s}{\lambda}$ [21]. From the above equations, it can be seen that if the matrices $\mathbf{K}_s$ and $\mathbf{K}_t$ are invertible, the weighting vectors α^s and α^t can be arbitrary vectors (with $\lambda = 1$) but result in a trivial solution. To overcome this problem, techniques like adding a additional regularization term or advancing reduced kernels [14] have been proposed. Adding a regularization term requires one to solve the following problem:

$$\mathbf{K}_t(\mathbf{K}_t + \kappa \mathbf{I})^{-1} \mathbf{K}_s \alpha = \lambda^2 (\mathbf{K}_s + \kappa \mathbf{I}) \alpha, \tag{7}$$

where $\alpha^t = \frac{(\mathbf{K}_t + \kappa \mathbf{I})^{-1} \mathbf{K}_s \alpha^s}{\lambda}$. However, this might not be preferable for large-scale problems due to the calculation/storage of a dense and large kernel matrix. In our work, we apply the reduced kernel technique [14] if a kernel version of CCA is needed. The use of reduced kernels allows us to solve eigenvalue decomposition problems efficiently while avoiding the over-fitting problem, as detailed in the following subsection.

B. Reduced Kernel for KCCA

In kernel methods, the reduced kernel technique favors large-scale problems in which the calculation of the full kernel matrix is computationally expensive. The key idea of the reduced kernel technique is to select a small portion of data and to generate a small rectangular kernel matrix for replacing the original full kernel matrix. Besides computationally more efficient, since one does not require the full kernel matrix, possible over-fitting or singularity problems can be avoided.

In the previous subsection, we have known that the projection vectors $\mathbf{u}^s$ and $\mathbf{u}^t$ can represented by the linear combination of the data $\mathbf{X}^s$ and $\mathbf{X}^t$ as in (4). We choose to approximate the weighting vectors (i.e., α^s or α^t) by using the linear combination of a subset of the original data, and thus reduced kernels can be calculated accordingly. Let $\tilde{\mathbf{X}}^s$ and $\tilde{\mathbf{X}}^t$ be subsets of $\mathbf{X}^s$ and $\mathbf{X}^t$, we have

$$\tilde{\mathbf{u}}^s = \tilde{\mathbf{X}}^s \tilde{\alpha}^s \text{ and } \tilde{\mathbf{u}}^t = \tilde{\mathbf{X}}^t \tilde{\alpha}^t, \tag{8}$$

where $\tilde{\alpha}^s \in \mathbb{R}^{\tilde{n}}$ and $\tilde{\alpha}^t \in \mathbb{R}^{\tilde{n}}$ are the weight coefficients for new/approximated linear combinations, respectively. As a result, $\tilde{\mathbf{u}}^s$ and $\tilde{\mathbf{u}}^t$ will be the approximated solutions of (5). To derive the reduced kernel matrices for KCCA, we apply Nyström approximation [14], [22], [23] for the full kernel matrix, i.e.,

$$k(\mathbf{X}^s, \mathbf{X}^s) \approx k(\mathbf{X}^s, \tilde{\mathbf{X}}^s) k(\tilde{\mathbf{X}}^s, \tilde{\mathbf{X}}^s)^{-1} k(\mathbf{X}^s, \tilde{\mathbf{X}}^s)^\top, \tag{9}$$

where $k(\mathbf{X}^s, \tilde{\mathbf{X}}^s) = \tilde{\mathbf{K}}_s$ is a reduced kernel matrix. Thus, we have

$$\begin{aligned} k(\mathbf{X}^s, \mathbf{X}^s)\alpha &\approx k(\mathbf{X}^s, \tilde{\mathbf{X}}^s) k(\tilde{\mathbf{X}}^s, \tilde{\mathbf{X}}^s)^{-1} k(\mathbf{X}^s, \tilde{\mathbf{X}}^s)^\top \alpha \\ &= k(\mathbf{X}^s, \tilde{\mathbf{X}}^s)\tilde{\alpha}, \end{aligned} \tag{10}$$

where $\tilde{\alpha}$ is an approximation of α via the reduced kernel technique. By the approximation, we can replace the full kernel matrices $k(\mathbf{X}^s, \mathbf{X}^s) \in \mathbb{R}^{n \times n}$ and $k(\mathbf{X}^t, \mathbf{X}^t) \in \mathbb{R}^{n \times n}$ by the reduced kernel matrices $k(\mathbf{X}^s, \tilde{\mathbf{X}}^s) \in \mathbb{R}^{n \times \tilde{n}}$ and $k(\mathbf{X}^t, \tilde{\mathbf{X}}^t) \in \mathbb{R}^{n \times \tilde{n}}$, respectively. The reduced kernel version for the KCCA (i.e., Rd KCCA) now solves the following problem:

$$\max_{\alpha^s, \alpha^t} \rho = \frac{\tilde{\alpha}^{s\top} \tilde{\mathbf{K}}_s^\top \tilde{\mathbf{K}}_t \tilde{\alpha}^t}{\sqrt{\tilde{\alpha}^{s\top} \tilde{\mathbf{K}}_s^\top \tilde{\mathbf{K}}_s \tilde{\alpha}^s} \sqrt{\tilde{\alpha}^{t\top} \tilde{\mathbf{K}}_t^\top \tilde{\mathbf{K}}_t \tilde{\alpha}^t}}, \tag{11}$$

where $\tilde{\mathbf{K}}_s = k(\mathbf{X}^s, \tilde{\mathbf{X}}^s)$ and $\tilde{\mathbf{K}}_t = k(\mathbf{X}^t, \tilde{\mathbf{X}}^t)$. For the simplicity of represention, we re-express (11) by

$$\max_{\alpha^s, \alpha^t} \rho = \frac{\tilde{\alpha}^{s\top} \Sigma_{\tilde{K}_{st}} \tilde{\alpha}^t}{\sqrt{\tilde{\alpha}^{s\top} \Sigma_{\tilde{K}_s} \tilde{\alpha}^s} \sqrt{\tilde{\alpha}^{t\top} \Sigma_{\tilde{K}_t} \tilde{\alpha}^t}}, \tag{12}$$

where $\Sigma_{\tilde{\mathbf{K}}_{st}} = \tilde{\mathbf{K}}_s^\top \tilde{\mathbf{K}}_t$, $\Sigma_{\tilde{K}_s} = \tilde{\mathbf{K}}_s^\top \tilde{\mathbf{K}}_s$, and $\Sigma_{\tilde{K}_t} = \tilde{\mathbf{K}}_t^\top \tilde{\mathbf{K}}_t$. The same as (2), we can solve (12) via

$$\Sigma_{\mathbf{K}_{st}} (\Sigma_{\mathbf{K}_{tt}})^{-1} \Sigma_{\mathbf{K}_{st}}^\top \tilde{\alpha}^s = \eta \Sigma_{\mathbf{K}_{ss}} \tilde{\alpha}^s. \tag{13}$$

The reduced kernel method reduce the computational cost from $O(n^3)$ to $O(\tilde{n}^3)$ (typically $n \ll \tilde{n}$). It has been shown that reduced kernel not only can avoid the over-fitting but also can approximate the solution of the full kernel matrix well.

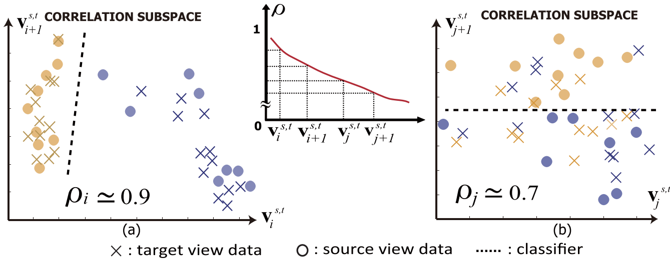


Fig. 3. Projecting source and target view instances from the IXMAS dataset into different correlation subspaces using CCA projection vectors with different ρ.

C. Parameter Selection for KCCA

In our work, we use Gaussian kernel functions for all KCCA or Rd KCCA, and thus we need to determine the parameter σ when calculating the Gaussian kernels. Since only *unlabeled* data pairs from both source and target domains are available during the derivation of the KCCA (or Rd KCCA) subspaces for domain adaptation, one cannot perform traditional cross-validation to select the parameter. To address this issue, we use k-mean clustering to divide these unlabeled data pairs into k clusters as cross-validation data (we simply set k as the number of classes to be recognized). In each validation fold, we take $k-1$ folds for generating the correlation subspace, and we evaluate the correlation coefficient of the projections of the remaining validation-fold data. This would indicate the capability of the associated KCCA or Rd KCCA model in describing cross-domain data in this subspace. Finally, we choose the kernel parameter resulting the highest correlation performance on these unlabeled data pairs.

IV. Correlation-Transfer SVM

A. Domain Transfer Ability of CCA

As discussed earlier, each dimension $\mathbf{v}_i^{s,t}$ in the derived CCA subspace is associated with a different correlation coefficient ρ_i. A higher ρ_i indicates a better correlation between data projected from different domains, which results in a better *domain transfer ability* for the associated dimension $\mathbf{v}_i^{s,t}$ in domain adaptation. From the above observations, it is clear that one should take domain transfer ability into consideration when designing classifiers in the CCA subspace.

Fig. 3 illustrates this issue by projecting source and target view data onto different 2D correlation subspaces, in which one subspace is associated with ($\mathbf{v}_i^{s,t}$ and $\mathbf{v}_{i+1}^{s,t}$) with higher $\rho \simeq 0.9$, and the other one is constructed by ($\mathbf{v}_j^{s,t}$ and $\mathbf{v}_{j+1}^{s,t}$) with smaller $\rho \simeq 0.7$. The dash lines represent linear classifiers learned from projected source view data (no labeled data in the target domain is available). From Fig. 3(a), we see that the location of projected source and target data with the same label are close to each other, since the two basis vectors correspond to larger ρ values. On the other hand, as shown in Fig. 3(b), the distributions of projected source and target view data are different due to a lower ρ. As a result, the classifier learned from projected source view data (i.e., the dash lines) cannot generalize well to the projected target view ones. In other words, poorer domain transfer ability will result in increased recognition error, even the classifier is well designed using the projected source view data.

To overcome such limitations for CCA during domain adaptation, we advocate the *adaptation* of the learning model based on the associated domain transfer ability. Since support vector machine (SVM) has been shown to be very effective in solving classification problems, we propose a new SVM formulation which particularly takes domain adaptation ability into account, so that the proposed model can be applied to address cross-view recognition.

B. The Proposed SVM Formulation

1) Classification in the CCA/KCCA Subspace: Generally, if the ith feature attribute exhibits better discrimination ability, the standard SVM would produce a larger magnitude for the corresponding model (i.e., a larger $|w_i|$). As discussed earlier, domain adaptation via CCA does not take the domain transfer ability into consideration when learning the classifiers in the correlation subspace. As a result, the recognition performance might be degraded. To address this problem, we introduce a correlation regularizer and propose a novel linear SVM formulation which integrates the domain transfer ability and class discrimination in a unified framework. Due to the introduction of such ability, the generalization of our SVM for transfer leaning will be significantly improved.

To incorporate domain adaptation ability into the SVM formulation, we first modify the standard SVM as follows:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i - \frac{1}{2}\mathbf{r}^\top \mathrm{Abs}(\mathbf{w}) \tag{14}$$
$$\text{s.t.}\ \ y_i(\langle \mathbf{w}, \mathbf{P}^{s\top}\mathbf{x}_i^s\rangle + b) + \xi_i \geq 1,\ \ \xi_i \geq 0,\ \ \forall(\mathbf{x}_i^s, y_i) \in \mathcal{D}_l^s,$$

where $\mathrm{Abs}(\mathbf{w}) \equiv [|w_1|, |w_2|, \ldots, |w_d|]$ and $\mathbf{r} \equiv [\rho_1, \ldots, \rho_d]$ is the correlation vector in which each element indicates the correlation coefficient of CCA for each projection dimension. Note that only labeled source domain data $\mathbf{x}_i^s \in \mathcal{D}_l^s$ is available for training (not target domain data), and y_i is the associated class label. Parameters C and ξ are penalty term and slack variables as in the standard SVM. We have $\mathbf{P}^{s\top}\mathbf{x}_i^s$ as the projection of source domain data $\mathbf{x}_i^s$ onto the correlation subspace $\mathcal{X}_c$.

From (14), it can be seen that the proposed term $\mathbf{r}^\top\mathrm{Abs}(\mathbf{w})$ is introduced for model adaptation based on CCA, and it is in terms of a similarity measure for $\mathbf{r}$ and $\mathbf{w}$. In practice, if a smaller correlation coefficient ρ_i is observed for the ith dimension of the CCA subspace, the above formulation would enforce the shrinkage of the corresponding $|w_i|$ and thus suppresses the learned SVM along that dimension. On the other hand, a larger ρ_i favors the contribution of the associated $|w_i|$ when minimizing (14).

Since it is not straightforward to solve the minimization problem in (14) with $\mathrm{Abs}(\mathbf{w})$, we seek the approximated solution by relaxing the original problem into the following form:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i - \frac{1}{2}(\mathbf{r}\odot\mathbf{r})^\top(\mathbf{w}\odot\mathbf{w}) \tag{15}$$
$$\text{s.t.}\ \ y_i(\langle \mathbf{w}, \mathbf{P}^{s\top}\mathbf{x}_i^s\rangle + b) + \xi_i \geq 1,\ \ \xi_i \geq 0,\ \ \forall(\mathbf{x}_i^s, y_i) \in \mathcal{D}_l^s,$$

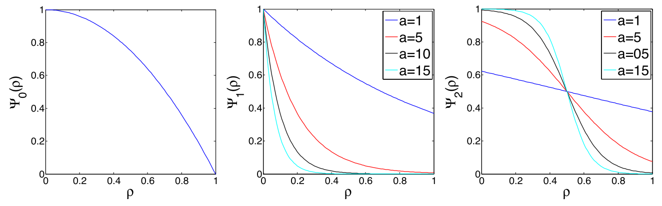


Fig. 4. Three types of weighting functions considered for our proposed method: (a) $\Psi_0(\rho) = 1 - \rho^2$, (b) $\Psi_1(\rho) = e^{-a\rho}$, and (c) $\Psi_2(\rho) = \frac{1}{1+e^{a(\rho-0.5)}}$.

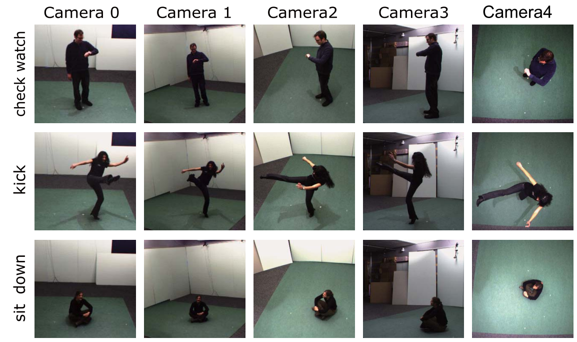


Fig. 5. Example actions of the IXMAS dataset. Each row represents an action at five different views.

where $\odot$ indicates the element-wise multiplication. We can further rewrite (15) as:

$$\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{d}(1-\rho_i^2)w_i^2 + C\sum_{i=1}^{N}\xi_i \tag{16}$$
$$\text{s.t.} \quad y_i(\langle \mathbf{w}, \mathbf{P}^{s\top}\mathbf{x}_i^s\rangle + b) + \xi_i \geq 1,\ \ \xi_i \geq 0,\ \ \forall(\mathbf{x}_i^s, y_i) \in \mathcal{D}_l^s.$$

We refer to (16) as our correlation-transfer SVM. Since we have the correlation coefficient $0 < \rho_i < 1$ in CCA, the convexity of the above optimization problem can be guaranteed. We note that, the implementation of our proposed SVM can be realized by modifying the software package of smooth support vector machine[1] [24], which applies the Newton-Armijo algorithm for solving SVM optimization problems. It can be seen that, depending on the derived correlation coefficients, the formulation in (16) is effectively weighting each component of the regularization term accordingly. As a result, this modified SVM automatically adapts the derived classification model $\mathbf{w}$ based on the domain transfer ability of CCA, and thus it exhibits better generalization in recognizing projected unseen test data in the correlation subspace (as confirmed by our experiments). The decision function for classifying unseen test data in target domain is shown as follows:

$$f(\mathbf{x}) = \text{sgn}\ (\langle \mathbf{w}, \mathbf{P}^{t\top}\mathbf{x}^t\rangle + b), \tag{17}$$

where $\mathbf{P}^t$ projects the input test data $\mathbf{x}^t$ from the target domain onto the correlation subspace $\mathcal{X}_c$, and sgn$(z) = 1$ if $z > 0$ and sgn$(z) =$ -1 if $z < 0$.

2) A Universal Form of the Correlation-Transfer SVM: In (14) or (16), we convert the standard SVM into an optimization problem, which takes the correlation coefficients of CCA into account when separating data between different classes in the resulting subspace. Our proposed SVM formulation thus assigns pre-determined weights for different coordinates in the CCA subspace. From (16), we see that each coordinate w_i is weighted by $1 - \rho_i^2$, and this weighting function can be illustrated in Fig. 4(a). Without loss of generality, we further reformulate our proposed SVM formulation as follows:

$$\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{d}\Psi(\rho_i)w_i^2 + C\sum_{i=1}^{N}\xi_i \tag{18}$$
$$\text{s.t.} \quad y_i(\langle \mathbf{w}, \mathbf{P}^{s\top}\mathbf{x}_i^s\rangle + b) + \xi_i \geq 1,\ \ \xi_i \geq 0,\ \ \forall(\mathbf{x}_i^s, y_i) \in \mathcal{D}_l^s,$$

where $\Psi(\rho)$ is the weighting function. In additional to the quadratic weighting function $\Psi_0(\rho) = 1 - \rho^2$, we consider exponential and sigmoid functions as two types of weighting functions for comparisons. For the exponential weighting function, it has the following form:

$$\Psi_1(\rho) = e^{-a\rho},$$

where ρ is the correlation coefficient and a is the parameter of $\Psi_1(\rho)$. The resulting exponential weighting functions with different parameters a are plotted in Fig. 4(b). In Ψ_1, we can see that the weight is decreasing remarkably when the associated correlation coefficient ρ drops. This type of weighting function will be of particular interest if only *few* high correlation projection vectors can be utilized. As for the sigmoid weighting function, it is determined as:

$$\Psi_2(\rho) = \frac{1}{1+e^{a\rho}},$$

where a is the parameter controlling the slope of $\Psi_2(\rho)$, and we plot this function with different parameters a in Fig. 4(c). Unlike prior quadratic or exponential weighting functions, the sigmoid one Ψ_2 exhibits additional flexibility in assigning weights to control the shrinkage of w. Later in our experiments, we will evaluate and choose proper weighting functions for achieving the best performance in each classification task.

V. Experiments

A. Datasets and Properties

In our experiments, we consider three datasets which exhibit different properties (and difficulty) in terms of domain adaptation. The first dataset is the *IXMAS multiview action dataset* [25], which has been widely used to address cross-view action recognition problems. This dataset contains 11 action categories, with 36 instances per class (and thus $n = 396$ in total). For this dataset, we follow the feature extraction settings as [7] did for comparisons, and thus the features in both source and target domains are of 1000 dimensions. When applying this dataset for verifying the effectiveness of our proposed framework, we have the scenario in which the features extracted from both domains are of the same type and $d_s = d_t = 1000 \gg n$.

[1]The SSVM toolbox is available at http://dmlab8.csie.ntust.edu.tw/ssvmtoolbox.html

	camera0					camera1					camera2					camera3					camera4				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
cam0			-			9.75	65.72	30.11	75.57	70.45	11.27	61.08	20.27	78.22	73.44	6.63	60.61	17.61	78.31	70.66	8.33	48.01	15.34	70.08	69.54
cam1	9.85	67.52	39.87	75.95	73.63			-			9.66	57.95	25.09	76.04	73.03	9.94	56.91	21.21	76.61	70.75	9.56	47.44	24.15	71.31	68.33
cam2	8.71	63.92	31.72	75.76	71.51	7.67	62.51	32.86	75.01	70.61			-			5.87	61.27	31.16	79.55	76.36	9.28	55.11	34.75	74.15	66.96
cam3	5.87	59.28	37.59	77.84	72.57	10.04	57.39	35.32	75.47	71.06	5.31	57.29	31.91	77.94	71.51			-			10.13	45.51	24.72	74.05	66.21
cam4	7.01	57.48	37.59	75.38	72.12	9.09	55.02	34.75	74.05	69.54	9.28	57.29	42.05	77.65	74.24	10.7	53.13	30.21	74.24	69.24			-		
Avg.	7.86	62.05	36.7	76.23	72.12	9.14	60.16	33.26	75.02	69.54	8.88	58.4	29.83	77.46	73.06	8.29	57.98	25.05	77.18	71.24	9.33	48.77	24.74	72.4	67.76

Fig. 6. Performance comparisons on the IXMAS dataset. Note that each row indicates the source view camera (for training), and each column is the target view camera for recognizing the action classes. We consider the methods of A: BoW without transfer learning [17], B: BoBW [7], C: linear CCA + linear SVM, D: linear CCA + our SVM (with $\Psi_1(5)$), and E: Rd KCCA + our SVM (with $\Psi_2(10)$).

The second dataset we consider is the *Multiple Features Dataset* from the UCI Machine Learning repository [26]. This dataset consists of features of handwritten digits from '0' to '9', and each is represented by six different features. With this dataset, our goal is to evaluate the recognition performance when adapting models across different feature spaces. It is worth noting that, we have d_s (and d_t) $\ll n$, which is very different from the *IXMAS multiview action dataset*. For this image recognition task, the use of this dataset allows us to assess the ability of our proposed framework for adapting information from one feature type to another, while both features of interest are extracted from the image data.

Finally, as the most challenging case, we consider the *Wikipedia dataset* [13] and address heterogeneous domain adaptation problems. While more detailed data pre-processing techniques and settings are described in the following subsections, we note that we will verify the effectiveness and robustness of our proposed method in solving cross-domain (i.e., image vs. text) classification tasks using this dataset.

B. IXMAS Multi-View Dataset

This dataset contains videos of eleven action classes. Each action video is performed three times by twelve actors, and the actions are synchronically captured by five cameras (i.e., five views/domains), as shown in Fig. 5. For fair comparisons with [7] and other approaches, we extract descriptors defined by [17] and describe each action video as a group of spatio-temporal cuboids (at most 200). For each video, these cuboids are quantized into $N = 1000$ visual words. As for data partition, we randomly choose two thirds of the cross-domain instances in each class as unlabeled data pairs for learning the CCA subspace. The remaining labeled data in the source domain are for training the proposed SVM, and those in the target domain are for testing. We repeat the above procedure ten times and report the average recognition performance.

To compare our performance with other approaches, we consider the methods of standard SVMs learned at the source view (i.e., using extracted BoW models without transfer learning), the bag-of-bilingual-words (BoBW) model proposed in [7], and CCA [2]. We note that, the above three approaches apply the standard SVM after deriving the feature representation for training/testing. The recognition results of different approaches are shown in Fig. 6.

From Fig. 6, we see that the method without transfer learning (i.e., columns A) achieved the poorest results as expected. While the BoBW model (columns B) and the approach of CCA (columns C) improved the performance

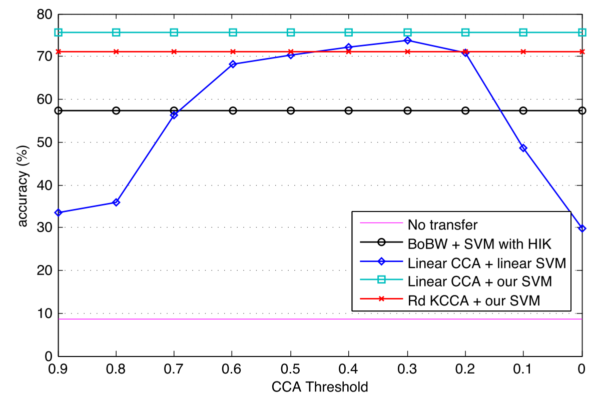


Fig. 7. Average recognition rates on the IXMAS dataset across different thresholds when deriving the CCA subspace. Note that our proposed method does not require to select such thresholds during domain adaptation.

by determining a shared representation for training/test, the integration of CCA with our proposed SVM (columns D) achieved the best performance. Comparing the results shown in columns B and D, although our SVM taking the correlation of the source and target view data was able to improve the recognition performance, it would be desirable to derive such correlation from a correlation-based transfer learning approach as our approach did. This explains why our approach combining CCA and imposing the resulting correlation coefficient into the proposed SVM formulation achieved the best recognition performance. It is worth noting that, comparing columns D and E, the use of nonlinear CCA for domain adaptation did not produce further improved recognition results. This is due to the fact that the action data are of 1000 dimensions, but only 396 instances are available for training (2/3 for deriving CCA, and 1/3 for training the CTSVM). For a high dimensional space with the low density of instances, linear models are expected to alleviate possible overfitting problems (i.e., the curse of dimensionality) during domain adaptation, and thus linear CCA with our proposed SVM achieved the the best recognition.

To evaluate the effectiveness and robustness of our proposed SVM, Fig. 7 further compares our results with those produced by CCA-based methods using predefined thresholds for determining the dimensions of the resulting subspace. We see that, for the case of linear CCA + linear SVM, while higher thresholds for the correlation coefficient resulted in lower dimensional subspace for domain adaptation, the use of such smaller number of dimensions did not exhibit sufficient domain transfer ability. On the other hand, if a lower threshold was applied, the use of more feature dimensions might

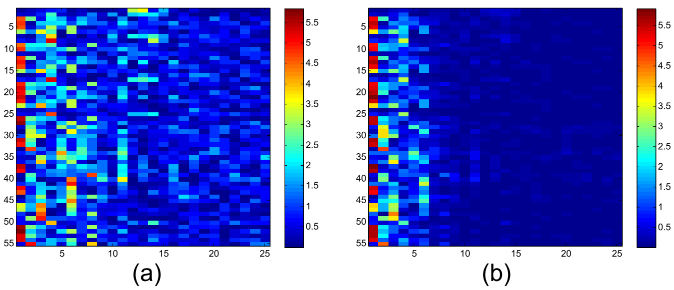


Fig. 8. Comparisons of the averaged $|w_i|$ values: (a) standard SVM and (b) our proposed SVM. The horizontal axis indicates the dimension index of the correlated subspace (arranged in a descending order using the associated correlation coefficients). The vertical axis shows the index of the derived one-vs-one SVMs. The recognition rates for these two approaches are 59.85% and 65.91%, respectively.

incorporate redundant or noisy information, which actually degraded the performance. Therefore, how to determine a proper threshold would be a challenging task for such standard domain adaptation approaches. From Fig. 7, it can be seen that our proposed SVM (with linear or nonlinear CCA) consistently achieved improved or satisfactory performance without the need to select a predetermined threshold.

Finally, we show the averaged value $|w_i|$ of each attribute in the standard and our SVM models in Fig. 8(a) and (b), respectively. From Fig. 8(a), we see that the standard SVM aims at separating data in the correlated subspace without considering the domain transfer ability, and thus we still observe prominent $|w_i|$ values at non-dominant feature dimensions (e.g., the 11th dimension). On the other hand, in Fig. 8(b), our proposed SVM suppressed the contributions of non-dominant feature dimensions in the correlated subspace, and thus resulted in larger $|w_i|$ values for dominant feature dimensions. The actual recognition rates for the two models were 59.85% and 65.91%, respectively. This observation verifies that our proposed SVM model with the ability of learning the domain transfer ability is preferable when solving cross-domain classification tasks.

C. Multiple Features Dataset

The Multiple Features Dataset [26] contains ten classes of handwritten numbers with 200 images per class. Thus, a total of 2,000 images are available for training and testing. These digits are represented in terms of six different types of features and dimensions, which are Fourier coefficients (76), profile correlations (216), Karhunen-Love coefficients (64), pixel average (240), Zernike moments (47) and morphological features (6) (the numbers in the parenthesis represent the dimensionality of each feature). In contrast to the IXMAS action dataset, the features of this dataset describe the images in different perspectives, which can be viewed as heterogeneous features with different dimensionalities. We consider the task of cross-domain classification as the domain adaptation problem, in which the source domain data are available and obtained from one of the feature spaces, while the target domain represent another feature space where the test data are presented. For this particular dataset, the features from source and target domains are of different types but extracted from the same images. This is different from the use of IXMAS dataset where the features are the same but obtained by images captured by different camera views. Therefore, the domain adaptation problem is expected to be more challenging than the prior task using the IXMAS dataset. It is worth noting that some feature types are with low dimensionality, which implies that the use of nonlinear domain transfer models might be preferable (as confirmed later).

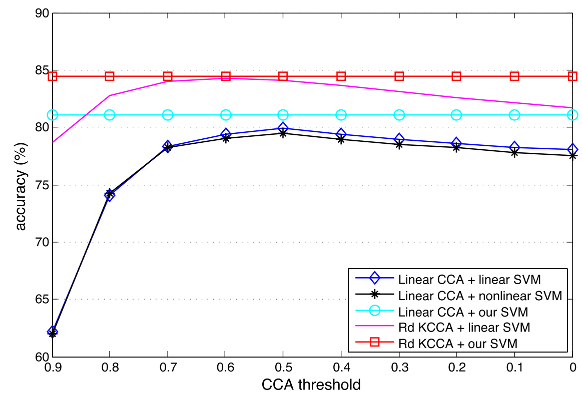


Fig. 9. Average recognition rates on the *Multiple Features* dataset across different thresholds when deriving the CCA subspace. Note that our proposed method does not require to select such thresholds during domain adaptation.

We shown the recognition results using different domain adaptation approaches in Fig. 9. It can be seen that, our proposed method achieved the best performance, and the use of Rd KCCA for domain adaptation consistently outperformed that of linear ones. We note that, while the dimension of the linear CCA subspace is bounded by the lower dimensional feature representation (from source or target domain), the above results verified that a nonlinear extension of CCA is more capable of efficiently searching an effective correlation subspace while performing the task of domain adaptation. Similar to our prior tests on the IXMAS dataset, we also observe that our method does not require the selection of CCA threshold when deriving the subspace. This provides additional robustness than other approaches. Finally, it is worth noting that the nonlinear extension of CCA with our proposed SVM outperformed the method of "linear CCA + nonlinear SVM" in this experiment. This observation implies that deriving a more effective feature space for domain adaptation is more crucial than designing a more complex classifier in a standard feature subspace.

D. Wikipedia Dataset

For the final part of our experiments, we consider a difficult heterogeneous domain adaptation task using the Wikipedia dataset [13], in which each data instance is with a text-image pair. In our experiment, we consider five subject categories of this dataset, which are art and architecture, biology, literature, sport, and warfare. This is because the visual data of these five categories are observed to exhibit lower intra-class variations (i.e., more relevant to the corresponding subjects). We select 100 instances from each cateogry for our experiments. For the textual representation, the features are derived from a latent Dirichlet allocation (LDA) model [27], which describes each textual instance by the probability of the topic assignment

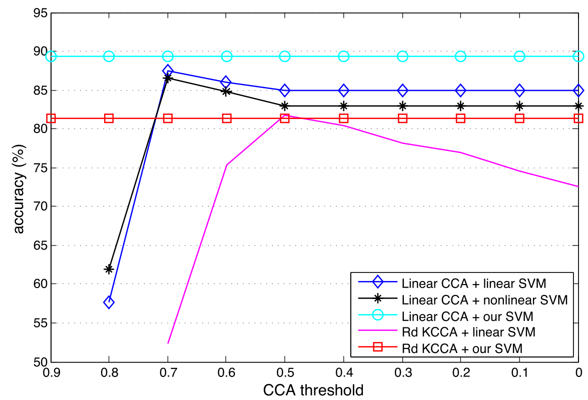


Fig. 10. Average recognition rates on *Wikipedia* dataset (image-to-text) across different thresholds when deriving the CCA subspace.

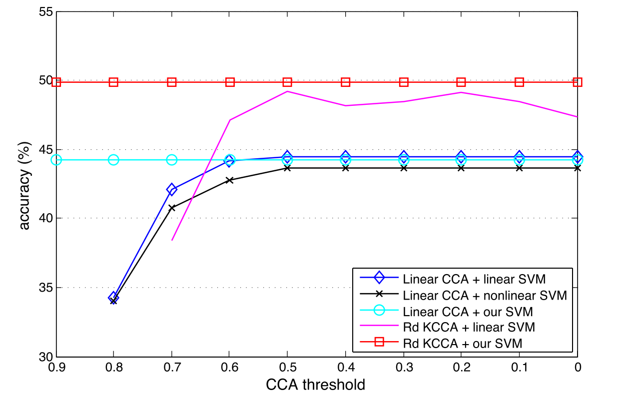
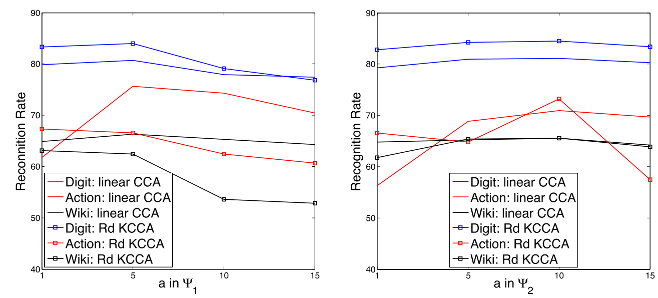


Fig. 11. Average recognition rates on *Wikipedia* dataset (text-to-image) across different thresholds when deriving the CCA subspace.



(a) (b)

Fig. 12. Comparisons of recognition rates using different a of (a) Ψ_1 and (b) Ψ_2 weight functions for each dataset. Note that Action, Digit, and Wiki stand for IXMAS Multi-View, UCI Multiple-Features, and Wikipedia datasets.

(and thus results in a five-dimensional vector for our case). On the other hand, the image data are represented by local descriptors of SIFT [28], and they are converted by a bag-of-words (BOW) model using the codebook constructed by k-means clustering (k is set to 128). As a result, the dimension of the image data is a 128-dimensional feature vector.

As for data partition for our implementation, we randomly choose two thirds of instances as unlabeled data, and the rest are labeled data for training purposes. Once the correlation subspace is derived, we train one-vs-one SVMs using projected source-domain data, and thus the projected target domain data can be recognized accordingly. We repeat the above setting for each combination of source-target pair, and report the average performance of twenty runs.

To demonstrate the effectiveness of our method, we perform single domain classification (i.e., text-to-text and image-to-image classification tasks) as baselines. The results are 95.40% and 56.20% for text-to-text and image-to-image tasks, respectively. For the cross-domain experiment, we first consider image-to-text recognition, and we show the results and comparisons in Fig. 10. By comparing with standard SVM and ours, we see that our proposed SVM for domain transfer clearly outperformed the standard one. We note that linear CCA with our proposed SVM achieved higher recognition rates than the reduced kernel CCA with our SVMs. Although the feature dimensions for text data is 5, they are extracted and summarized from dimensional word space. Such features effectively contain representative information for performing recognition, which is verified by a very high text-only recognition rate of 95.4% using only $d = 5$. Thus, this is consistent with observations made in other learning tasks with high-dimensional data and low sample density (e.g., cross-view action recognition in our pervious experiment), i.e., linear models are sufficient for discriminating such data between different classes while alleviating possible over-fitting problems. On the other hand, Fig. 11 shows the results for text-to-image classification problems. From this figure, it can be seen that the reduced kernel CCA with our SVM outperformed other methods. Similar to observations made in Multiple Feature Dataset, the use of nonlinear models could introduce additional capability in representing and separating lower dimensional data. Nevertheless, from our cross-domain classification tests using this dataset, the effectiveness and robustness using our proposed SVM models for addressing heterogeneous domain adaptation tasks can be successfully confirmed.

E. Parameter Selection for Weight Functions

As discussed in Section IV-B, we need to determine the parameter a for our proposed correlation-transfer SVM. Unlike the standard SVM classifier which one can perform cross-validation on training data (of the same domain) for choosing the associated parameters, our proposed SVM is only accessible to labeled data projected from the source domain, while the test data projected from the target domain will remain unseen. In other words, it would not be desirable to perform cross-validation only on projected source-domain data due to potential overfitting problems. Although we set this parameter empirically in our experiments, we found the recognition performance of our correlation-transfer SVM is not sensitive to the choice of parameter a for the weighting functions. As shown in Fig. 12(a) and (b), we observe that $a = 5$ and 10 achieved the best performance for weight functions Ψ_1 and Ψ_2, respectively, while other a choices still produced improved recognition rates when comparing to the standard SVM classifier. This indicates that our correlation-transfer SVM, which takes domain adaptation ability of the derived subspace into consideration, is effective in utilizing the representative feature dimensions and thus exhibits excellent capabilities in recognition. We also note that, in all our experiments, we selected the best parameters for all other methods and reported their performance for fair comparisons.

VI. Conclusion

In this paper, we proposed a novel approach for heterogeneous domain adaptation and classification. By exploring the correlation subspace derived by CCA using unlabeled data pairs of source and target view data, we presented a novel SVM formulation with a correlation regularizer. The proposed correlation-transfer SVM takes the domain transfer ability into consideration when designing the classifier in the correlation subspace. As a result, only projected and labeled training data from the source view are required when designing the classifier in the resulting subspace (i.e., no training data at the target view is needed). Experimental results on three cross-domain datasets confirmed the use of our proposed framework for improved recognition, and we verified that our approach outperformed state-of-the-art transfer learning algorithms which did not take such domain transfer ability into consideration.

References

[1] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[2] J. Blitzer, D. Foster, and S. Kakade, "Domain adaptation with coupled subspaces," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 173–181.

[3] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 213–226.

[4] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. IEEE ICCV*, Nov. 2011, pp. 999–1006.

[5] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2012, pp. 2066–2073.

[6] A. Farhadi and M. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 154–166.

[7] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2011, pp. 3209–3216.

[8] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2012, pp. 2855–2862.

[9] T. van Kasteren, G. Englebienne, and B. Kröse, "Transferring knowledge of activity recognition across sensor networks," in *Proc. 8th Int. Conf. Pervas. Comput.*, 2010, pp. 283–300.

[10] S. Pan, J. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. 23rd AAAI Nat. Conf. Artif. Intell.*, 2008, pp. 677–682.

[11] S. Pan, I. Tsang, J. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, pp. 1187–1192.

[12] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[13] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[14] Y.-J. Lee and S.-Y. Huang, "Reduced support vector machines: A statistical theory," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 1–13, Jan. 2007.

[15] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," in *Proc. Adv. NIPS*, 2007, pp. 601–608.

[16] M. Sugiyama, S. Nakajima, H. Kashima, P. Von Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Proc. Adv. NIPS*, 2007, pp. 1433–1440.

[17] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop VS-PETS*, Oct. 2005, pp. 65–72.

[18] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2, pp. 107–123, 2005.

[19] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1785–1792.

[20] L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for heterogeneous domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 711–718.

[21] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2364, 2004.

[22] C. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Proc. Adv. NIPS*, 2001, pp. 682–688.

[23] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a gram matrix for improved kernel-based learning," *J. Mach. Learn. Res.*, vol. 6, pp. 2153–2175, Dec. 2005.

[24] Y. Lee and O. Mangasarian, "SSVM: A smooth support vector machine for classification," *Comput. Optim. Appl.*, vol. 20, no. 1, pp. 5–22, 2001.

[25] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 249–257, 2006.

[26] A. Frank and A. Asuncion. (2010) *UCI Machine Learning Repository* [Online]. Available: http://archive.ics.uci.edu/ml

[27] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

Yi-Ren Yeh received the M.S. and Ph.D. degrees from the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taiwan, in 2006 and 2010, respectively. From 2008 to 2009, he was a Visiting Scholar with CyLab, Carnegie Mellon University, Pittsburgh, PA, USA. He was a Post-Doctoral Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan, and the Intel-NTU Connected Context Computing Center, National Taiwan University, Taipei, from 2010 to 2013. He is currently an Assistant Professor with the Department of Applied Mathematics, Chinese Culture University, Taipei. His research interests include machine learning, data mining, optimization, numerical methods, and pattern recognition.

Chun-Hao Huang was born in Tainan, Taiwan, in 1985. He received the B.S. degree in electrical engineering and the M.S. degree in computer and communication engineering from the National Cheng Kung University, Tainan, in 2008 and 2010, respectively. After military duty, he was with the Research Center for Information Technology Innovation, Academia Sinica, in 2011. He is currently pursuing the Ph.D. degree with Technische University Mnchen. His research interests include 2-D/3-D conversion, human motion capture, and computer vision related issue.

Yu-Chiang Frank Wang received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 2001. From 2001 to 2002, he was a Research Assistant with the National Health Research Institutes, Taiwan. He received the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2004 and 2009, respectively.

Dr. Wang joined the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan, in 2009, where he is currently a Tenure-Track Associate Research Fellow. He leads the Multimedia and Machine Learning Laboratory, CITI, and involved in the fields of computer vision, pattern recognition, machine learning, and image processing. He serves as an Organizing or Program Committee Member at multiple international conferences or activities, and several of his papers were nominated for the Best Paper Awards at related conferences. In 2013, he is selected as the Outstanding Young Researcher by the National Science Council of Taiwan.